平成25年度 修士研究論文

特徴量学習による
名詞と形容詞の同時認識

電気通信大学大学院 情報理工学研究科
総合情報学専攻 メディア情報学コース
1230035　　氏名 小原 侑也
主任指導教員　　柳井 啓司 准教授
指導教員　　尾内 理紀夫 教授

平成26年1月30日

概 要

近年のデジタルカメラの台頭、画像投稿サイトやSNSの出現により、テキストやタグが関連づけられた画像がWeb上に多く存在している。このように多量な画像の中から複数のタグを基にAND検索を行うと、本来求めていないような画像も検索結果に現れる。こうした想定外のノイズとなるような画像を除去するためには、画像に写った物体やシーンを認識し、検索ワードに対して適している画像のみを分類する必要がある。しかし、名詞と形容詞の組み合わせは莫大な数になり、全ての組み合わせに対して認識を行うにはコストが大きくなりすぎる。そこで、画像認識によってノイズが除去できるような特徴量を持つ組み合わせか否かを判断する手法を提案する。

本研究では、まず10種類の動物の名詞と各名詞と共起の高い10形容詞の組み合わせに対して画像を収集し、収集した画像から画像特徴量を抽出した。そして抽出した特徴量の分布をエントロピーの手法を用いて数値化し、画像の同時認識によって精度が高くなるような分布が狭くなる特徴量を抽出可能なクラスを求め、どのようなクラスで独特な特徴量が抽出できたかの考察を行った。画像特徴量は、色や形状以外の抽象的な形容詞にも対応するため、特徴量学習の手法を用いて抽出した。

その結果、beautiful+lionやcute+catのようなクラスで特徴量の分布が減少することを確認することができ、特徴量学習の手法を用いたことで、抽象的な形容詞に関しても、他と区別できるような適切な特徴量が抽出できた事を確認できた。

目 次

第1章

はじめに

1.1　背景

近年、デジタルカメラの台頭、またインターネット利用の浸透に伴い、Web 上に存在する画像枚数が爆発的に増加している。また、Flickr 等の画像投稿サイトや、Twitter、Facebook 等の SNS の出現により、画像には多くのタグやテキストが関連付けされている。

そのような画像の増加に伴い、画像検索の有用性、また難易度も高くなっている。画像検索の際に、複数のタグやテキストに基づく AND 検索を行うと、期待していた画像と異なる画像が結果として提示されることがしばしば起こる。これは、タグやテキストが不適切に付けられていることや、タグ間の関係が考慮されていないこと、また主観的な要素が入ることで、検索結果として視覚的な分布が広い、見た目がバラバラな画像集合が提示されるためだと考えられる。例えば、“カッコイイ犬” で画像検索を行うと、図 1.1 のような画像が提示されるが、これらに何らかの視覚的な一貫性があるかは自明でない。

しかし、“カッコイイ犬” のような特定の語の組み合わせに関して、人間が検索結果として求める画像に視覚的な一貫性があるならば、その特性に基づいた画像の分類が可能となるはずである。視覚的な特性に基づいた分類の手法として、画像中の詳細な内容を認識する画像認識の手法がある。画像認識の手法においては、認識対象となるクラスの数だけ画像の分類器を用意し、それぞれの出力結果を比較して最も尤度の高いクラスへ分類する。

タグの AND 検索のような複数の語の組み合わせにおいて認識を行うには、全ての組み合わせ、すなわち名詞と形容詞の積に相当する莫大な数に対して分類器を作成して各分類器の出力を得る必要がある。そのため、検索するごとに毎回分類

処理を行おうとすると、多大な時間が掛かる。語の組み合わせの中には本来あり得ないような組み合わせや、画像の視覚的な分布がバラバラで、組み合わせても認識精度が向上しない組み合わせも存在する。このような組み合わせに対して認識を行っても、検索結果の質の向上は期待できないため、認識対象から外すことで、認識処理にかかる時間を低減させられる。そこで、認識を行うべき組み合わせか否かを判断する指標が必要になる。

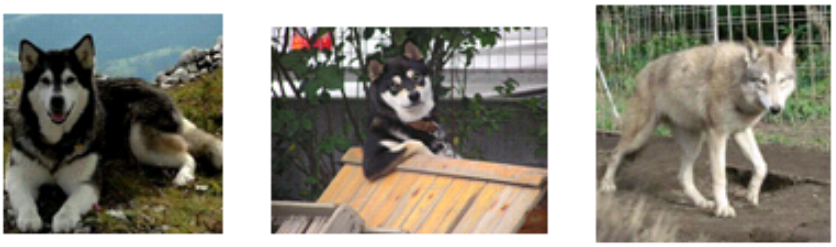

図 1.1: カッコイイ犬で収集した画像

1.2　目的

本研究の目的は、認識を行った際に高い認識精度が見込める組み合わせか否かを判断することである。高い認識精度が見込めるか否かは、特定の語と語を組み合わせて検索を行った際に収集される画像集合に、視覚的な観点における特定の分布が見受けられるかという観点から判断する。視覚的な分布は画像から抽出される特徴量を基に計算する。分布の偏った特徴量を持つ画像集合は他の集合と区別しやすいと考えられるため、このようなクラスに関しては名詞と形容詞の組に対する画像の同時認識を用いてノイズ除去を行うことで、画像検索時における結果の質を向上させることができると思われる。その一方で分布の偏らない特徴量を持つクラスにおいては同時認識を行う必要性が低いと考えられるため、認識を行うクラス数を低減させ、コストがなるべく掛からないような処理を行えるのではないかと考えている。

同時認識とは、２つの単語が同時に示す物体・場面を学習・認識する事である。すなわち、“可愛い”＋“犬” のような組み合わせに該当する画像を認識することを指し示す。人間がある一点を指しながら、それを複数単語によって表現する時、多くの場合には形容詞＋名詞の形式で表現を行う。

本研究では形容詞と名詞の組み合わせに着目し、分布の計算を行う。ここで問題となるのは、形容詞には主観を伴うものなど、指し示す範囲の広い単語の存在である。そのため、形状や色に特化した特徴量では画像の表現に限界がある。その対策として、本研究においては、手動ではなく、画像の画素値から自動で特徴量を学習・抽出を行う、特徴量学習の仕組みを用いることとした。

1.3　本論文の構成

本論文の構成は以下のようになっている。

1 章　はじめに
　本研究の背景、目的を記載する。
2 章　関連研究
　関連研究を紹介する。
3 章　提案手法の概要
　本研究の概要を記載する。
4 章　提案手法の説明
　本研究の詳細と利用手法について説明する。
5 章　実験
　本研究における実験と実験結果を記載する。
6 章　考察
　実験結果に対する考察を記載する。
7 章　おわりに
　まとめと今後の課題を記載する。

第2章

関連研究

本研究では、特徴量学習を使用して抽出した画像特徴を用いて、名詞と形容詞の間にある視覚的な関連性の分析と、両者の組み合わせに対応する物体・場面の同時認識を行っている。形容詞は属性の一形態と考えられるため、ここでは属性を用いた画像認識の研究と、特徴量学習を用いた画像認識、物体検出の研究を紹介する。

2.1　属性を用いた研究

近年では、属性を用いて画像を認識しようとする研究が多数行われている。属性とは、画像内に写っている物体そのものの名前ではなく、赤や縞模様のような物体の色やテクスチャ、また、コップの取っ手のようにその有無で物体の見た目が変化するような物体の構成パーツなどを表す語のことを示す。

属性を用いた研究としては、例えば、K.Duan らの研究 [1] が挙げられる。Duan らは、条件付き確率場の考え方を基に、鳥や蝶の画像から、特定の属性に関する局所的な領域の検出を行っている。検出例を図 2.1 に示す。例えば図 2.1 の右下の写真を見れば、赤い目とお腹、さらに黒い尾を持つ鳥が写っており、かつ目やお腹がどこに写っているのかも一目で分かるようになっている。

また、Y.Han らの研究 [2] では、色やテクスチャといった視覚的な属性を対象としている。Han らは属性がラベルとして付けられた画像データセットを用いて、未知画像に含まれるであろう属性を予測する研究を行っている。属性予測の流れを図 2.2 に示す。[2] では、物体クラスと属性、属性と属性の間の関係を行列やグラフ構造で表現し、そこから未知画像の属性の予測を行う。

図 2.1: 論文 [1] の属性領域の検出結果

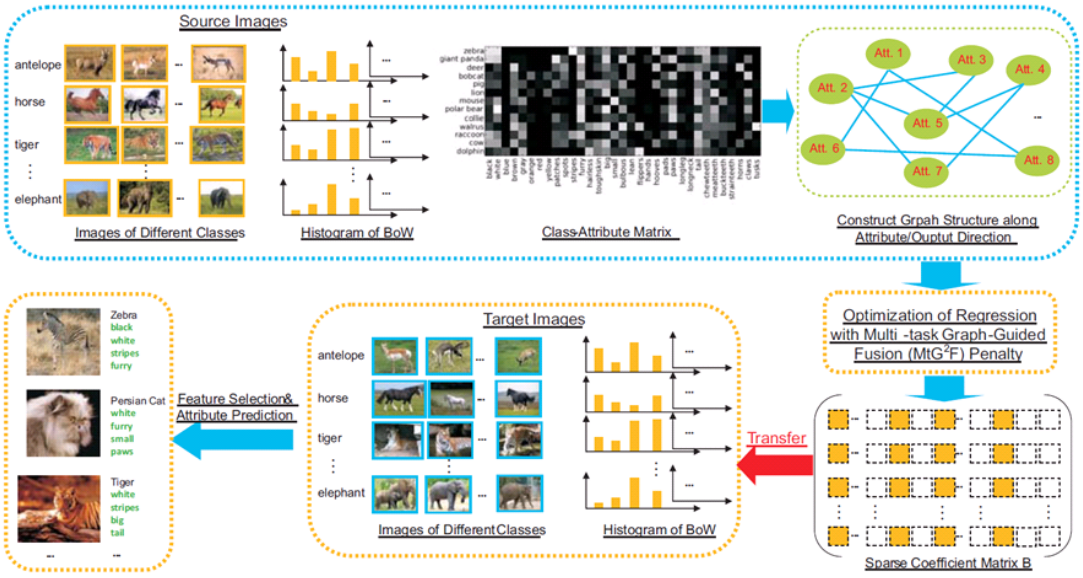

図 2.2: 論文 [2] の属性予測の手法

上記で挙げた研究では画像に含まれる属性の予測や属性を含む領域の検出を目的としているが、本研究では属性としての形容詞が既に付与された画像の視覚的な分布を計算することを目的とする。

2.2　特徴量学習を用いた研究

近年、既存の人手によって作成される特徴量では表現しきれない画像の特徴を表すために、画像のピクセル値から、その画像を表現するのに最適な表現を学習する、特徴量学習の手法を用いた研究が行われている。

特徴量学習を用いた研究として、P.Wu らの研究 [3] が挙げられる。Wu らの研究の目的は、オンライン学習が可能な類似画像検索のためのアルゴリズムの提案である。Wu らの特徴量抽出アルゴリズムを図 2.3 に示す。Wu らは画像に適した表現が可能な特徴量抽出を行うために、Stacked Denoising Autoencoder を利用した Deep Learning の手法を用いている。Wu らは特徴量の学習として異なるデータセットの画像から複数の特徴量を用いて計算を行うため、、それぞれを別の Deep Learning によって計算し、それらをマルチ様式に対応するよう統合する手法によって特徴量の学習を行っている。また、類似画像検索の結果を図 2.4 に示す。各クエリ画像に対して、Wu らの手法を用いた類似画像検索を行った最下段の結果が最もクエリ画像と類似しているのが見て取れる。

この研究と本研究の違いとしては、クエリ画像との類似ではなく、画像集合を対象として、その画像集合が属するクラスの視覚的な類似性に着目している点がある。

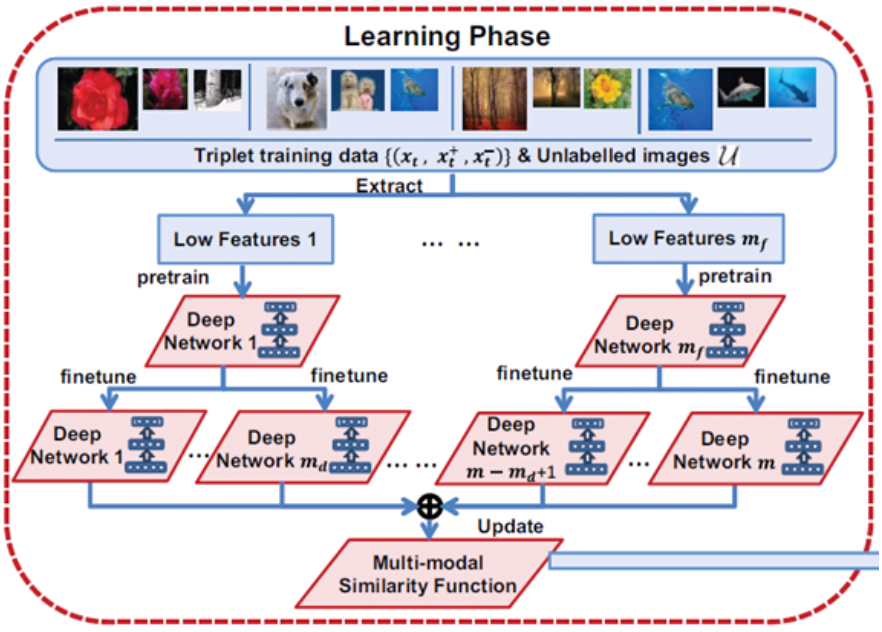

図 2.3: 論文 [3] のアルゴリズム

図 2.4: 論文 [3] の類似画像検索結果

また、A.Krizhevsky らの研究 [4] では、Convolutional Neural Network を利用した Deep Learning を用いて画像認識を行っている。図 2.5 は ILSVRC-2010 test set に対する認識のエラー率に関して自動特徴量学習と手製の特徴量と比較したものであるが、従来の人手で作成した特徴量を用いた画像認識精度を上回る、高い認識精度を誇っていることが分かる。抽出の流れを図 2.6 に、認識結果を図 2.7 に示す。また、論文 [4] において提案された手法を用いた画像認識プログラムは [5] にて公開されている。 Convolutional Neural Network による学習では教師ありの学習データが必要であり、学習画像の枚数も多く必要とする。一方、本研究にて用いた k-means learning の手法では、教師なしの学習データでよいという違いがある。そのため、k-means learning を用いることは、収集可能な画像枚数が少なくなる場合や、学習画像に確実なラベルが付いていない場合などに適している。

本研究では、既存の手製の特徴量では表現しづらい主観的な形容詞にも対応するため、特徴量の自動学習手法として k-means learning の手法を用いる。k-means

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

図 2.5: [4] による特徴量と手製の特徴量によるエラー率の比較

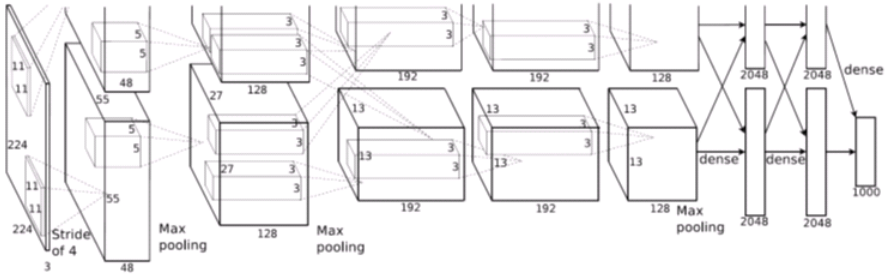

図 2.6: 論文 [4] の特徴抽出の流れ

図 2.7: 論文 [4] の画像認識結果

learning は A.Coates らによって論文 [6] にて提案された手法である。k-means learning ではまず、学習画像からランダムに切り出したパッチを k-means によってクラスタリングして代表的なパッチを定めて辞書とする。そして、特徴抽出をしたい画像から等間隔で切り出したパッチと最も類似している辞書内のパッチに投票を行うことで画像を数値化する。

2.3　単語の視覚性の評価

次に、本研究の先行研究となる論文の紹介を行う。ここで挙げるのは、柳井らの研究 [7]、秋間らの研究 [8] および、川久保らの研究 [9] である。

柳井らの研究では、単語概念における視覚性を定量化する方法としてエントロピーを提案し、150 個の形容詞に着目して視覚的関連性について言及している。本論文内で使用するエントロピーによる単語の視覚性を数値化する方法、また、エントロピーの計算方法は、柳井らの研究によって提案された手法を用いる。

秋間らの研究では、概念間の距離関係や上下関係から概念間の階層構造を持ったデータベースを構築している。上下関係を求める際にエントロピーを使用し、画像の分布を計算している。また、画像に付与されているタグ情報も使用している。

川久保らの研究では、単語概念における視覚性と地理的分布を求めている。この研究では、視覚性を求めるために、領域分割やエントロピー計算を用いて、名詞や形容詞といった概念クラスの画像分布の計算を行っている。

本研究においては、特定の名詞と形容詞のタグが両方とも付いている画像を対象として、組み合わせごとのエントロピーを計算する。そして、得られたエントロピーを基に、その名詞と形容詞の組み合わせたクラスに含まれる画像を表す特徴量が他のクラスと区別可能な偏った分布を持っているか否かを判断する。

また、我々の今までの研究としては [10] が挙げられる。[10] では、名詞と形容詞のどちらも固定して決め、それらの名詞と形容詞を組み合わせたクラスに含まれる画像を対象として、視覚的な分布を計算した。しかし、このような語の組み合わせを行うと、人が検索しようと考え得ないような語の組み合わせが発生し、分布計算の必要性が低まる事が考えられる。本研究においては、固定の名詞に対して、その名詞と共起の高い形容詞を選択することで、より人が検索に用いやすいような組み合わせを対象として計算を行う。さらに、[10] では人手による特徴量である Color-SIFT を用いているが、色や形状ではない主観的な語には対応しづらいことから、このような語に対応できるよう、自動特徴量学習の手法を用いて抽出した特徴量を用いるように変更した。

第3章

本手法の概要

本研究では、名詞と形容詞を組み合わせて画像検索を行った際に収集される画像集合に視覚的な観点から見た偏りがあるか否かを判断する。

実験においては固定の名詞に形容詞を追加することとするが、形容詞も固定のものにすると、通常は同時には使われない組み合わせが出現する。例えば、“circle+beach” のような組み合わせが挙げられる、図 3.1 に “circle+beach” のタグ検索によって収集した画像を示すが、浜辺が写った画像が少なく、当然ばらつきが多いことも見て取れる。このような場合、そもそも同時認識の必要性が低い、画像枚数が極端に少ない等の事が考えられるため、本実験では、固定の名詞と共起の高い形容詞を選出し、その組み合わせに対応する画像を収集している。

図 3.1: circle+beach でタグ検索を行った結果

また、人が表現に用いる形容詞には、主観的なものや質感的なものなど、既存の人手により作成する特徴量では数値として表現しづらい語も含まれるため、そのような語への対応が期待できる視覚的な特徴表現方法を取り入れる。

特定の名詞と形容詞が付与された画像によって学習を行い同時認識の精度が高まるには、収集される画像集合の視覚的な関連性が高い、すなわち画像集合から

抽出される特徴量の分布が偏っていることが重要であると考えられる。そこで、本手法ではエントロピーの概念を用いて特徴量分布を数値化する。

本研究における実験の実行手順を図 3.2 に示す。

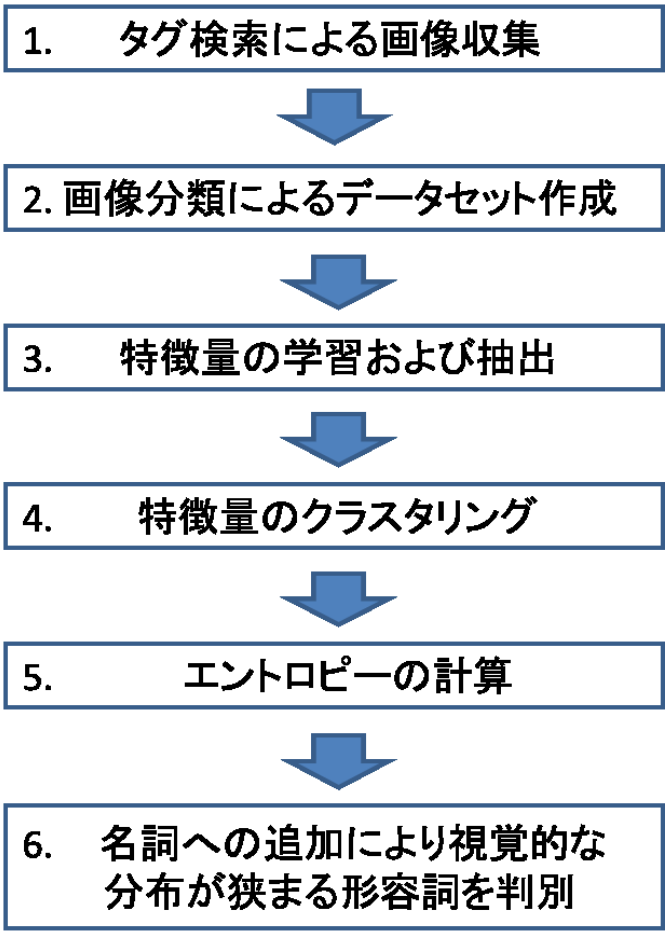

図 3.2: 提案手法の流れ

ステップ 1 では、本論文の分析に使用する画像を、形容詞と名詞の AND 検索を用いて収集する。

ステップ 2 では、ステップ 1 で収集した画像から、人手によって、本来分析対象としている動物が写っている画像のみを各クラス 100 枚分類する。

ステップ 3 では、分類した画像を対象とし、自動特徴量学習の手法である k-means learning を用いて、画像に適した特徴量の学習および抽出を行う。

ステップ 4 では、pLSA を用いて画像から抽出した特徴ベクトルをクラスタリングする。ここでは、エントロピーの計算に必要な画像内での隠れトピックの同時確率 $P(z|d)$ を求める。

ステップ 5 では、名詞と形容詞を組み合わせたクラス概念が音の画像分布を、エントロピーを計算することによって求める。エントロピーは画像分布が狭まる、すなわち画像特徴量の分布が狭くなると減少し、分布が広まると増大する。よって、このステップで求めたエントロピーが小さくなるほど、その名詞と形容詞を組み合わせたクラスと他のクラスとを区別できる特徴量が抽出できていると判断できる。

ステップ 6 では、ステップ 5 で求めたエントロピーを基にして、各名詞と形容詞を組み合わせて収集した画像集合から抽出した特徴量に視覚的な偏りが存在するか否かを判断し、そのような画像集合を用いて同時認識を行うべきかを考察する。

第4章

本手法の説明

本章では、本研究において用いられた手法の詳細について説明を行う。

4.1　画像収集

本研究では人手によってタグを付けられた画像を用いる。そのため、タグ付け可能な画像投稿サイトである Flickr より、Flickr API を用いて画像収集を行った。画像収集の初期段階として、名詞と形容詞の組み合わせ 1 組ごとに 1,000 枚の画像を収集した。収集の際には、名詞をクエリタグ 1、形容詞をクエリタグ 2 として AND 検索を行い、両方のタグが画像と紐付けされた画像を検索対象とした。また、同一投稿者が、ほぼ同一な画像を連続して投稿することも大いに考えられるため、同一投稿者からの取得には 2 枚の制限を設けた。

4.1.1　Flickr API

本小節では、画像収集の際に利用した Flickr API について述べる。Flickr API とは、画像投稿サイトである Flickr によって提供された Web API である。Flickr API の利用には、Yahoo.com の ID を取得し、Flickr にユーザ登録することで取得可能となる API Key が必要となる。

Flickr API を利用して画像検索を行うには、Flickr.photos.search メソッドを用いる。Flickr.photos.search メソッドを利用する際には、パラメータ変更によって画像検索の条件や取得データ内容の変更が可能となる。今回利用したパラメータを表 4.1 に示す。なお、本研究では、tag_mode を all、sort を interestingness-desc として、画像検索および収集を行った。

表 4.1: flickr.photos.search メソッドのパラメータの一例

パラメータ名		説明
api_key	必須	取得した API Key
method	必須	メソッド (写真検索は flickr.photos.search)
content_type	任意	検索画像の種類を指定する 1:写真のみ、2:スクリーンショットのみ、 3:写真およびスクリーンショット以外の画像のみ、 4:1 と 2 に含まれる画像、5:2 と 3 に含まれる画像、 6:1 と 3 に含まれる画像、7:すべての画像
tags	任意	タグ (テキストタグ) を指定する
page	任意	ページ番号を指定する
per_page	任意	1 ページあたりに検索する画像数を指定する
tag_mode	任意	tags で複数のタグを指定した場合の検索方法を指定する (any:OR 検索、all:AND 検索)
sort	任意	検索結果のソート方法を指定する (date-posted-asc(desc):アップロード日時の古い (新しい) 順、 date-taken-asc(desc):撮影日時の古い (新しい) 順、 interestingness-desc(asc):人気が高い (低い) 順、 relevance:関連度が高い順)
extras	任意	追加出力項目を指定する (license:ライセンス種別、date_upload:アップロード日時、 owner_name:投稿者名、icon_server:アイコンサーバー、 original_format:アップロード時のフォーマット、 last_update:更新日時、geo:ジオタグ (緯度・経度))

flickr.photos.search メソッドの API リクエスト URL 例

http://www.flickr.com/services/method=flickr.photos.search&api_key=*****
&tags=car,red&tag_mode=all&content_type=1&sort=interestingness-desc
&per_page=500&page=1&extras=date_taken

4.2　画像分類

前節で述べたような手法に基づいて収集した画像には、本来目的とするような画像とは異なる、タグとして使用した名詞や形容詞と関連が低い画像も収集されてしまう可能性がある。そこで本研究では、手動で本来目的としておらずノイズとなるような画像の除去を行った。画像分類の工程では、名詞のクエリタグに相当する動物がなるべく大きく写っている画像をポジティブ画像とし、名詞と形容

詞を組み合わせたクラスごとに 100 枚ずつ収集した。

4.3　特徴量学習

本研究では、画像の数値的な表現である画像特徴量として、A.Coates ら [6] によって手案された k-means learning の手法を用いて学習・抽出を行った特徴量を用いる。k-means learning の手法では、従来提案されてきた人手によって設計される特徴量とは異なり、画像のピクセル値から、その画像の表現に最適な特徴量を自動で学習している。

Coates らによって提案された特徴量学習の手順は以下のようになっている。

- 教師無し学習画像セットからランダムにパッチを切り出す。
- 切り出したパッチに前処理を加える。
- 教師無し学習アルゴリズムによる特徴量マッピングの学習を行う。

上記の手順で行われている前処理とは、正規化と白色化を示している。正規化とは、各ピクセルのピクセル値から全体のピクセル値の平均を引き、標準偏差で割る処理を指している。この処理を行うことで、平均を 0、分散を 1 とすることができるため、明度とコントラストに関して、正規化を行う処理となっている。白色化とは [11] で提案された、ベクトルの共分散行列を単位行列とする変換であり、Zero-phase whitening と呼ばれている。学習時においては、白色かをランダムに切り出したパッチ行列に対して適用している。

次に、特徴量抽出の手順を以下に示す。

- 画像から等間隔にパッチを切り出す。
- 切り出したパッチに対して正規化と白色化による前処理を施す。
- 各パッチに対して符号化の処理を行う。
- 画像を 2×2 に分割し、領域ごとに符号化された値の和を計算する。
- それぞれの領域の和を連結し、特徴ベクトルとして表現する。

上記の符号化とは、以下の式を適用する処理である。

$$f(x) = [f_1(x), f_2(x)), \cdots, f_K(x))] \tag{4.1}$$

$$f_k(x) = max(0, \mu(z) - z_k)) \tag{4.2}$$

$$z_k = ||x - c^k||_2 \tag{4.3}$$

ここで c^k とは学習時に求めた k 番目のクラスタ中心であり、$\mu(z)$ は全クラスタ中心との距離の平均値である。以上で説明した特徴抽出の処理の流れを図 4.1 に示す。

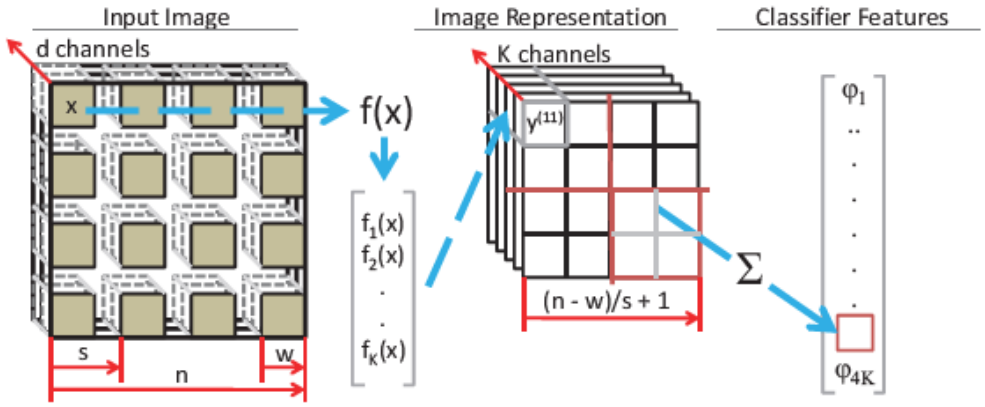

図 4.1: 特徴量抽出の流れ（論文 [11] より引用）

4.4　特徴量分布の計算

特徴量分布を計算するために、pLSA を利用して特徴ベクトルのクラスタリングを行った。

4.4.1　pLSA

pLSA(probabilistic Latent Semantic Analysis) は、T. Hofmann によって提案された、テキストコーパス内のトピック検出を行う、統計的言語処理のためのモデルである [12]。pLSA のグラフィカルモデルを図 4.2 に示す。

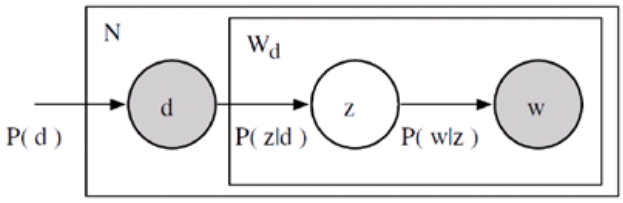

図 4.2: pLSA のグラフィカルモデル

pLSA の計算は次のように行う。まず、$d_i(i = 1, 2, \cdots, I)$ を文書、$w_j(j = 1, 2, \cdots, J)$ を単語、$z_k(k = 1, 2, \cdots, K)$ を隠れトピックとする。隠れトピックとは、文書内における単語の生成に関するトピック変数である。次に、文書 d_i と単語 w_j の生起は独立であると考え、文書 d_i と単語 w_j の同時確率 $P(d_i, w_j)$ を以下のように表す。

$$P(d_i, w_j) = \sum_{k=1}^{K} P(d_i|z_k)P(w_j|z_k)P(z_k) \tag{4.4}$$

また、文書 d_i 内において、単語 w_j が生成される確率 $P(w_j|d_i)$ は、隠れトピック z_k を用いて以下のように表すことができる。

$$P(w_j|z_k) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{4.5}$$

そして、文書 d_i 内での単語 w_j の出現回数を $n(d_i, w_j)$ と表すと、データの対数尤度は次のように表すことができる。

$$L = \sum_{i=1}^{I} \sum_{j=1}^{J} n(d_i, w_j) log(d_j, w_j) \tag{4.6}$$

この対数尤度を最大にするような $P(z_k)$、$P(d_i|z_k)$、$P(w_j|z_k)$ を EM アルゴリズムを用いた最尤推定によって求める。

pLSA は統計的言語処理のモデルであるが、文書を画像、単語を局所特徴ベクトルに置き換えて考えることによって、画像認識においても利用されている。

4.5　エントロピー

エントロピーは pLSA によって求められた確率を用いて計算を行った。エントロピーは特徴ベクトルの分布が広くなると増大し、狭くなると減少する。そのため、エントロピーの大小が、名詞と形容詞の組み合わせであるクラス概念に属する画像分布の広さを表す。図 4.3 は名詞 flower に形容詞 red を付けた場合に、画像の視覚的な差異が flower 画像よりも red flower 画像の方が少なくなることから、エントロピーが減少している様子を表している。以上のことから、すなわちエントロピーを計算することが、組み合わせごとの視覚的関連性を求める事に繋がると言える。

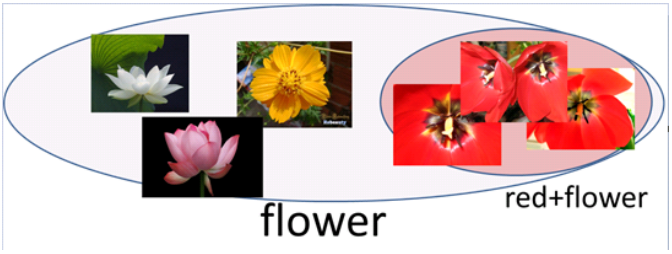

図 4.3: 形容詞付与によるエントロピー減少の様子

エントロピーの計算方法は以下のようになっている。pLSA を用いて求めた $P(z_k|d_i)$ を用いて計算を行う。各隠れトピック z_k に対して、

$$P(z_k|w_j) = \frac{\sum_{i=1}^{I} P(z_k|d_i)}{|I|} \tag{4.7}$$

を求める。次に、先ほど求めた $P(z_k|w_j)$ を用いて、各画像 d_i に対して、

$$H(P) = -\sum_{k=1}^{K} P(z_k|w_j) log(P(z_k|w_j)) \tag{4.8}$$

を求める。この計算によって求まる $H(P)$ がエントロピーである。

第5章

実験

本章では、本研究における実験の内容に関しての詳細および実験結果を示す。

5.1　画像収集

名詞と形容詞の組み合わせに対応した画像データセット作成のため、Flickr API を用いて Flickr より画像データを収集した。今回の実験においては、まず動物 10 種類を名詞として定め、それらの動物名と共起しやすい形容詞を組み合わせた “名詞＋形容詞” クラスを 10 クラスずつ用意した。各名詞に対する形容詞の決定手順は以下のようになっている。

- 決定した名詞がタグとしてつけられた画像 5000 枚分のタグ情報を収集
- 各タグの出現回数を計算
- 頻出タグ上位の語の内、形容詞としての意味を持つ語を 10 単語選択

上記のようにして名詞 10 単語それぞれに形容詞 10 単語を組み合わせた合計 100 組を最終的なデータセットのクラスとして採用した。ここで決定したクラスは表 5.1 に示す。

そして、クラスごとに各名詞と形容詞のタグが共に付与されている 1000 枚の画像を収集した。また、同一投稿者からの画像には類似した画像が含まれることも多く、画像の分布に極端な影響を与えることが考えられるため、同一投稿者からの画像は 2 枚までのみ収集するという制限を設けた。この段階で収集した画像の例を図に示す。

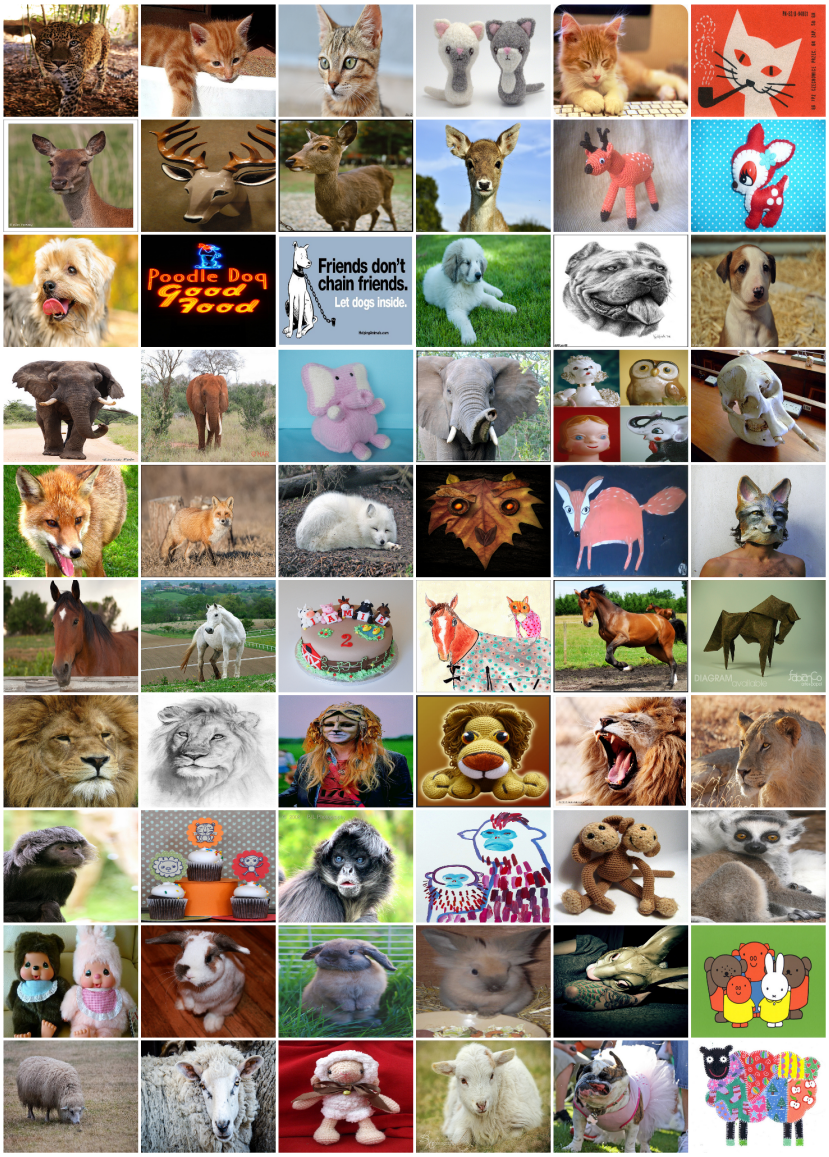

図 5.1: 各名詞に形容詞 “animal” を追加して集めた画像例

表 5.1: 名詞＋形容詞クラス一覧

cat	animal	cute	pet	white	black
	big	beautiful	furry	funny	sweet
deer	animal	forest	sunrise	red	green
	wild	white	cute	light	fallow
dog	pet	cute	animal	white	black
	green	light	funny	adorable	happy
elephant	animal	wild	cute	black	white
	green	forest	national	sunset	bush
fox	animal	red	wild	cute	white
	urban	green	national	forest	arctic
horse	animal	white	black	sunset	beautiful
	green	light	red	western	wild
lion	animal	big	wild	male	cute
	white	beautiful	black	green	female
monkey	animal	cute	mono	wild	black
	funny	forest	golden	young	small
rabbit	cute	animal	adorable	white	pet
	sweet	green	fluffy	funny	furry
sheep	green	animal	rural	white	light
	farming	cute	sunset	black	square

5.2　画像分類

前節のようにしてタグ情報のみを利用して画像収集を行うと、本来目的としている画像とは異なる画像や、クエリタグと関連の低い画像も含まれる。そのため今回の実験では、収集した画像から、クエリタグに選択した動物が写っている画像を、名詞と形容詞を組み合わせたクラスごとに 100 枚ずつ分類する過程を設け、形容詞付き動物画像データセットを作成した。本分類によって選択された画像と排除した画像の例を図 5.2、5.3 に示す。

図 5.2: データセットとして使用する画像例

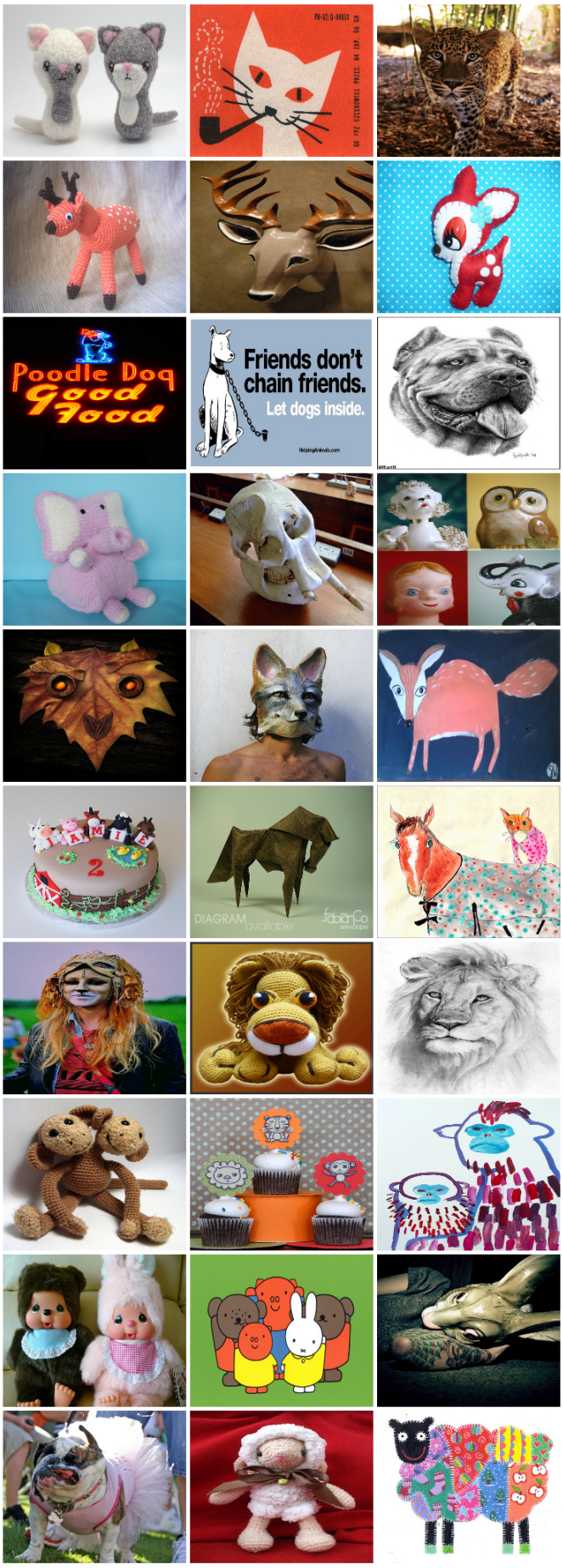
Poodle Dog
Good
Food
Friends don't
chain friends.
Let dogs inside.

図 5.3: 排除した画像例

5.3　特徴量学習

前節、前々節のようにして作成したデータセットより、特徴量の学習および抽出を行った。画像から切り出すパッチのサイズは 16 × 16 とし、1 ピクセルずつスライドさせながら等間隔にパッチの切り出しを行った。また、この過程によって抽出される特徴量の次元数は 1000 になるように設定した。

5.4　特徴量分布測定

前節にて求めた特徴量に対して pLSA を用いたクラスタリングを行い、特徴量分布を計算した。この際のクラスタリングにおけるクラスタ数は 300 とした。また、pLSA を利用して計算した同時確率 $P(z|d)$ を用いて、エントロピーの計算を行い、各クラスの視覚的な分布の広がり具合を数値化した。

5.5　比較実験

本研究では自動特徴量学習によって抽出した特徴量を用いている。そこで、比較のために人手による特徴量として Color-SIFT を用いて、同様の実験を行った。特徴ベクトルとして表現するために、画像から Color-SIFT を抽出し、BoF としてベクトル化した。BoF を求めるために、まず、コードブックを作成した。コードブックの作成には、タグによる制限を加えずランダムに選択した 10000 枚の画像を使用した。この 10000 枚の画像に対して Grid による特徴点決定を行い、Color-SIFT 特徴量を抽出し、全特徴量からランダムに 100 万個の特徴量を選択した。この 100 万個の特徴量に対して、k-means 法を用いて k=1000 のクラスタリングを行った。そのため、本実験にて使用したコードブックのコードブックサイズは 1000 となっている。次に、データセットとして用意した画像に対しても同様に Color-SIFT 特徴を抽出した。そして先ほどのコードブックを用い、各画像から 1000 次元の BoF を作成した。

5.6　実験結果

5.6.1　k-means learning による特徴量の結果

自動特徴量学習の手法として k-means learning を用いて特徴量を抽出し、分布の計算を行った結果を図 A.1 から図 A.19 に示す。また、各名詞に対して、エントロピーがより小さい順に形容詞を並べた結果を表 5.2 および 5.3 に示す。表 5.2、5.3 では、主観に影響されやすい形容詞として beautiful と cute を赤色で、色の形容詞として green、black および white を青色で提示している。また、形容詞の下には、そのクラスのエントロピーの値を記述している。

表 5.2: エントロピーが小さい順に形容詞を並べた表 (k-means learning(1))

cat	beautiful	animal	white	cute	furry
	2.09405	2.09737	2.09913	2.10228	2.10592
	black	sweet	funny	pet	big
	2.10891	2.10925	2.10953	2.11036	2.11207
deer	white	cute	red	forest	light
	1.91901	1.91978	1.91982	1.92118	1.92133
	wild	green	fallow	sunrise	animal
	1.92328	1.92877	1.93784	1.93874	1.94887
dog	green	happy	adorable	black	animal
	2.18667	2.19172	2.19325	2.19450	2.19470
	light	cute	pet	white	funny
	2.19664	2.19713	2.19913	2.20049	2.20200
elephant	cute	forest	national	animal	green
	2.11519	2.11731	2.11831	2.11900	2.11948
	wild	sunset	bush	white	black
	2.11949	2.12209	2.12971	2.13127	2.13404
fox	animal	white	cute	green	urban
	2.07640	2.07788	2.07877	2.08785	2.08987
	forest	wild	red	national	arctic
	2.09158	2.09172	2.09420	2.09711	2.10121

表 5.3: エントロピーが小さい順に形容詞を並べた表 (k-means learning(2))

horse	light	western	animal	beautiful	green
	1.91977	1.92251	1.93998	1.94105	1.94361
	wild	white	sunset	red	black
	1.94455	1.94928	1.95033	1.95093	1.96153
lion	beautiful	male	black	cute	green
	2.14552	2.15625	2.15983	2.16074	2.16401
	big	female	white	wild	animal
	2.16605	2.16678	2.17649	2.17923	2.17971
monkey	funny	small	mono	wild	black
	2.13425	2.14409	2.14843	2.15015	2.15467
	golden	cute	young	forest	animal
	2.15500	2.15734	2.15736	2.15858	2.15955
rabbit	pet	fluffy	white	furry	green
	1.99700	2.00017	2.00247	2.00382	2.00415
	cute	sweet	animal	adorable	funny
	2.00497	2.00601	2.00646	2.01445	2.01557
sheep	black	green	sunset	animal	cute
	2.08246	2.09686	2.09897	2.09959	2.10124
	white	rural	farming	light	square
	2.10397	2.10673	2.10876	2.11212	2.11594

5.6.2　Color-SIFT 特徴量による結果

自動特徴量学習の手法との比較として、手製の特徴量である Color-SIFT 特徴量を抽出し、分布の計算を行った結果を図 A.2 から図 A.20 に示す。また、各名詞に対して、エントロピーがより小さい順に形容詞を並べた結果を表 5.4 および 5.5 に示す。

表 5.4: エントロピーが小さい順に形容詞を並べた表 (Color-SIFT(1))

cat	big	sweet	white	pet	animal
	2.21309	2.21665	2.22498	2.22501	2.23148
	funny	black	beautiful	furry	cute
	2.24173	2.24454	2.24509	2.25367	2.26059
deer	fallow	light	green	forest	red
	2.16807	2.20560	2.20810	2.21437	2.21606
	cute	wild	sunrise	animal	white
	2.21667	2.22168	2.22639	2.23354	2.23797
dog	green	funny	pet	white	adorable
	2.12841	2.15245	2.17976	2.19108	2.19826
	happy	light	cute	animal	black
	2.20049	2.20793	2.21081	2.21397	2.21443
elephant	black	white	cute	animal	green
	1.87809	1.93215	2.10427	2.11055	2.11772
	forest	wild	bush	national	sunset
	2.11893	2.11955	2.12774	2.13350	2.17649
fox	green	urban	forest	national	wild
	2.09891	2.10343	2.15316	2.15811	2.17856
	red	cute	animal	white	arctic
	2.18894	2.20305	2.21813	2.23288	2.23522

表 5.5: エントロピーが小さい順に形容詞を並べた表 (Color-SIFT(2))

horse	green	beautiful	wild	red	animal
	2.08290	2.14063	2.15108	2.15197	2.15933
	black	white	western	light	sunset
	2.16929	2.17966	2.17979	2.20531	2.22851
lion	black	animal	beautiful	cute	big
	2.23017	2.23901	2.24057	2.24122	2.24480
	male	female	wild	white	green
	2.24509	2.24515	2.24737	2.25041	2.25126
monkey	funny	mono	young	forest	cute
	2.15798	2.16438	2.16885	2.17119	2.19074
	golden	animal	small	black	wild
	2.19174	2.19279	2.19331	2.19687	2.20703
rabbit	green	furry	funny	fluffy	pet
	2.18209	2.23030	2.24416	2.24507	2.25624
	animal	sweet	white	cute	adorable
	2.25746	2.25772	2.25842	2.26709	2.27397
sheep	square	green	farming	rural	animal
	2.03474	2.12475	2.12764	2.12961	2.13357
	black	white	light	cute	sunset
	2.13449	2.13567	2.14692	2.15678	2.16696

第6章

考察

本節では前節の実験および実験結果より、名詞と形容詞の視覚的な分布の偏りについて考察を行う。

6.1　選択した形容詞について

今回の実験における形容詞は名詞によって異なっている。そこで始めにどのような形容詞が含まれているのかを検討する。表 5.1 を参考に形容詞を見てみると、大きく 4 種類に分類できるように思われる。1 つ目は black や green のような色を表す形容詞、2 つ目は beautiful や adorable のような主観に影響されやすい形容詞、3 つ目は forest や western のような場面を表す形容詞、4 つ目は big や pet など名詞を直接修飾しやすい形容詞である。

6.2　名詞と形容詞の視覚的な分布

表 5.2、5.3 を参考に、k-means learning を用いて抽出した特徴量に関するエントロピーの値がより小さくなる、すなわち特徴量分布が狭く視覚的な偏りを持つクラスに着目する。

最初にエントロピーを小さくする上位 5 形容詞全体を眺めてみる。このとき、上位に入る形容詞の種類に大きな偏りはなくバラけている。

次に各名詞に対して最もエントロピーを小さくした形容詞に焦点を当てる。すると、cat、lion に対しての beautiful や、elephant に対しての cute のような人間の主観に影響されやすいような形容詞が挙がっている。先述のクラスに含まれる画像を図 6.1 に示す。図 6.1 を見ると、一貫性はあまり無いようにも感じられるが、様々な

人が cute や beautiful であると感じる画像を多数集め、特徴量学習により特徴抽出を行うと、beautiful や cute に特有な分布を示すことができた。これにより、主観的な語を適切に表すような特徴抽出が可能になっていると考えられる。また、5 位

図 6.1: beautiful+cat、beautiful+lion、cute+elephant に含まれる画像例

までを見ると、cat、deer、fox、lion、sheep に cute が、dog には happy や adorable がランクインしている。この中で、図 6.2 に示すような cute+dog や cute+cat といったクラスには、子猫や子狐の画像が多く含まれる結果となっている。cat や dog に他の形容詞を付与したクラスにおいては、見られない独特な偏りが生じているため、このような特徴を抽出した結果エントロピーが減少したのではないかと考えられる。

図 6.2: cute+cat、cute+fox に含まれる画像例

次に、色に関係する形容詞に着目する。エントロピーが小さくなる形容詞の上

位に入っているクラスには、white+deer や black+sheep が挙げられる。この 2 クラスに関しては、一般的に羊や鹿と言われて思い浮かべる色とは異なる色である。しかし、先述のような白色の鹿や黒色の羊は実際に存在しており、図 6.3 に示すようにそのような色を持った動物の画像が収集できている。そのため、sheep や deer が付く他のクラスにおいては、白い毛皮を持つ羊や赤茶色の鹿が多く、黒色の羊や白色の鹿は含まれていても少数派である。このように色に関しては、他のクラスと大きく異なるような色を持つ方が、白い羊のように一般的に思い浮かべるような色の形容詞を動物の名詞に付与するより、画像の視覚的な差異が減り、エントロピーが減少する特徴量が得られたと考えられる。

図 6.3: black+sheep、white+deer に含まれる画像例

6.3　Color-SIFT 特徴量との比較

本節では、本手法で用いた特徴量学習の手法と、手製による特徴量を用いた際の、特徴量分布の違いについて考察する。表 5.2、5.3 と表 5.4、5.5 の各動物に関してエントロピーを小さくするような形容詞を比較する。

k-means learning の手法を用いた場合では、前節の考察でも示したように、エントロピー低減率上位の形容詞には beautiful のような語が多く含まれていた。それに対して Color-SIFT を用いた場合には、エントロピーを最も小さくする形容詞として、dog、fox、horse、rabbit に対する green や elephant、lion に対する black のように、色を意味する形容詞が多くなっていることが見て取れる。色の中でも green は特に顕著な結果となっており、形容詞として green を含む全名詞に関して、全名詞に関して 5 位以内に収まっている。green を含むクラスとして、クラス、クラス、クラス、クラスの画像を図に示す。図を見て分かるように、背景が芝生のような緑

色に限定されている事、動物だけが画面全体にアップで写っているような構図が減っている事が見て取れる。k-means learning も Color-SIFT 同様、色情報を用いており、エントロピーを減少させる形容詞の表にも green 等の形容詞が入っているものの、Color-SIFT を用いた方が、色の情報に影響を受けやすいのではないかと考えられる結果となった。

図 6.4: green+fox、green+horse、green+rabbit に含まれる画像例

第7章

おわりに

7.1　まとめ

本論文では、名詞と形容詞を組み合わせたクラスに含まれる画像を対象として、その画像集合が狭い分布を持つクラスであるか否かを判定し、どのようなクラスで分布が狭くなるのかを考察した。なお、画像の数値的な表現として画像特徴量表現を用いた。本研究で使用した特徴量は k-means learning と呼ばれる自動特徴量学習の手法で抽出した特徴量である。特徴量学習を用いた理由としては、既存の手製の特徴量では、色や形状への対応力が強い一方で、cute などの主観的な形容詞に対応しづらいという性質を補うためである。

実験を行った結果として、k-means learning を用いて抽出した特徴量を用いると、beautiful や cute のような形容詞を組み合わせることでエントロピーが小さくなる事例を多く確認することができた。これは、人間が主観的な判断から形容詞タグを付与した画像集合から適切な特徴量を抽出でき、その画像集合に偏った分布があると判断することが可能になったと思われる。また、色についても、black+sheep や white+deer のように、他の多くのクラスに含まれる同一名の動物と比較して特徴的な見た目を持つクラスに関してもエントロピーの減少が見受けられた。このことから、色に関しても、他と視覚的に異なるような画像集合から独特な特徴量を抽出できた。その一方で、画像に含まれる多くの領域の色が変化するような green+rabbit 等のクラスに関しては Color-SIFT を用いる方が、より独特な特徴量が抽出されるという結果も得た。

本研究の結果として、既存の手製の特徴量で分布が小さくならなかった主観的な形容詞に関して、特徴量学習の手法を用いることで独特な特徴量を得ることができた。また、他のクラスと比較して独特な色を持つ動物のクラスの分布が下がっ

たことから、組み合わせの可否に関して、判断の指標の一つに使えるのではないかと考えられる。その一方で、人間が見た時に偏りがあるような色の形容詞ではColor-SIFT の方が分布を狭めていたことから、特徴量の選択や組み合わせなどの課題が残る結果となった。

7.2　今後の課題

今後の課題として、まず一つ、特徴量選択の手法を用いることが考えられる。今回の結果として、特徴量学習の手法を用いることにより、今まで課題となっていた主観的な語を表す特徴量を抽出できたが、その一方で、やはり色や形状に関しては手製の特徴量に分があるクラスも存在した。そこで予め複数の特徴量を抽出しておき、それらを選択したり組み合わせたりすることで、より名詞と形容詞の組み合わせに対応した画像に適した特徴量の獲得と、より正確な組み合わせるべきか否かの判断を行えるのではないかと考えられる。

また、本研究において画像データセットの作成、特に動物でない画像を除去する画像分類のコストが大きかったと感じられる。そこで、クラスタリングによって予め動物らしい画像を選別して分類画像枚数を減らす方法や、クラウドソーシングサービスを利用して分類を行う方法を取り入れる事も、将来的な課題である。

参考文献

[1] K. Duan, B. In, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2012.

[2] Y. Han, F. Wu, X. Lu, Q. Tian, Y. Zhuang, and J. Luo. Correlated attribute transfer with multi-task graph-guided fusion. 2012.

[3] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and A. C. Miao. Online multimodal deep similarity learning with application to image retrieval. 2013.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 1106–1114, 2012.

[5] cuda-convnet. *http://www.code.google.com/p/cuda-convnet/*.

[6] A. Coates, Andrew Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

[7] 柳井啓司, Kobus Barnard. 一般物体認識のための単語概念の視覚性の分析. 情報処理学会論文誌: コンピュータビジョン・イメージメディア, Vol. 48, No. SIG10 (CVIM17), pp. 88–97, 2007.

[8] 秋間雄太, 川久保秀敏, 柳井啓司. Folksonomy を用いた画像特徴とタグ共起に基づく画像オントロジーの自動構築. 電子情報通信学会論文誌. D, 情報・システム, Vol. 94, No. 8, pp. 1248–1259, 2011.

[9] 川久保秀敏, 柳井啓司. 単語概念の視覚性と地理的分布の関係性の分析. 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 8, pp. 1417–1428, 2010.

[10] Y. Kohara and K. Yanai. Visual analysis of tag co-occurrence on nouns and adjectives. In *Proc. of International Conference on Multimedia Modeling*, 2012.

[11] A. Hyvärinen and E. Oja. Independent component analysis: algorithms andapplicaitons. *Neural networks*, Vol. 13, No. 4, pp. 411–430, 2000.

[12] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, Vol. 43, pp. 177–196, 2001.

付 録A

エントロピーの計算結果

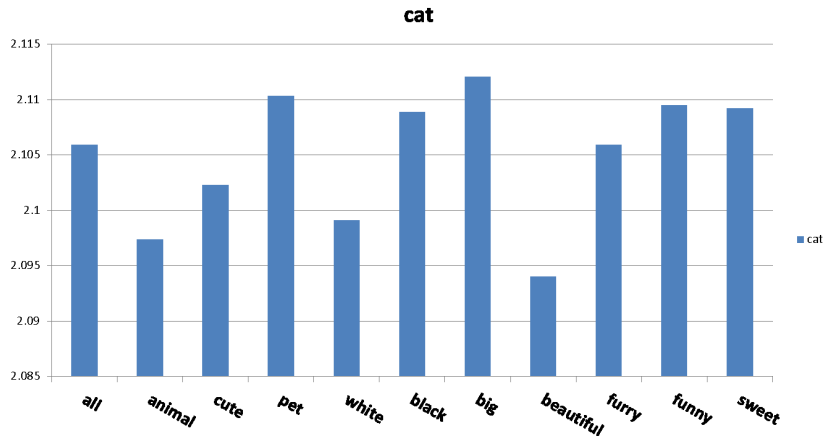

図 A.1: エントロピーの計算結果 (kmeans,cat)

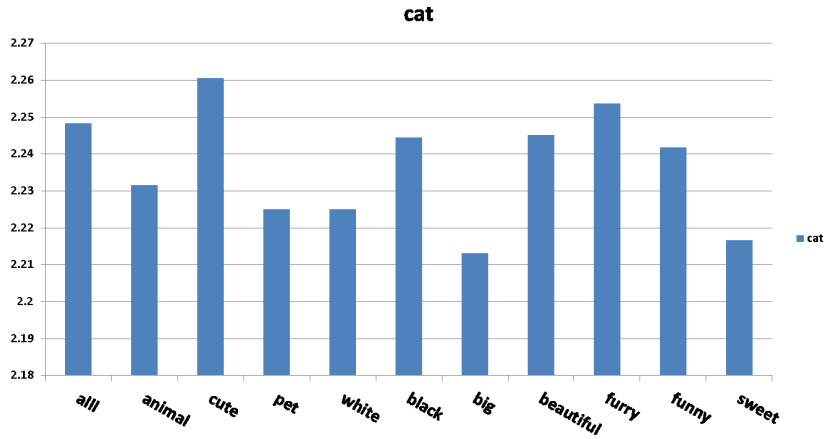

図 A.2: エントロピーの計算結果 (csift,cat)

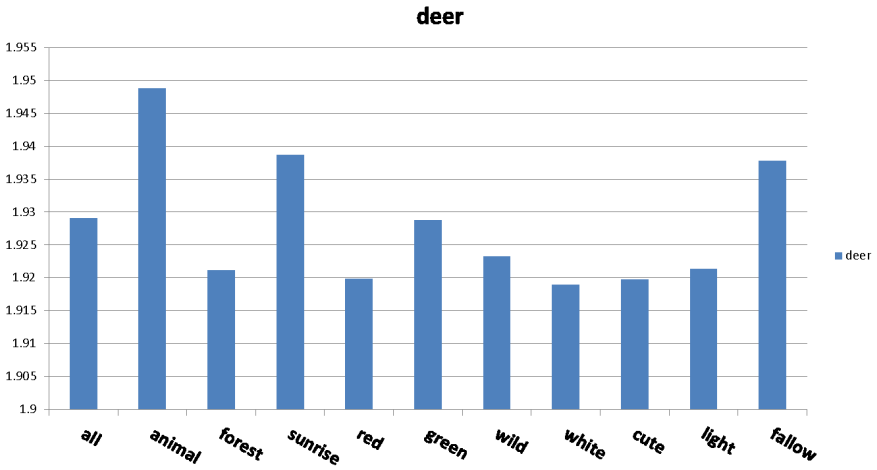

図 A.3: エントロピーの計算結果 (kmeans,deer)

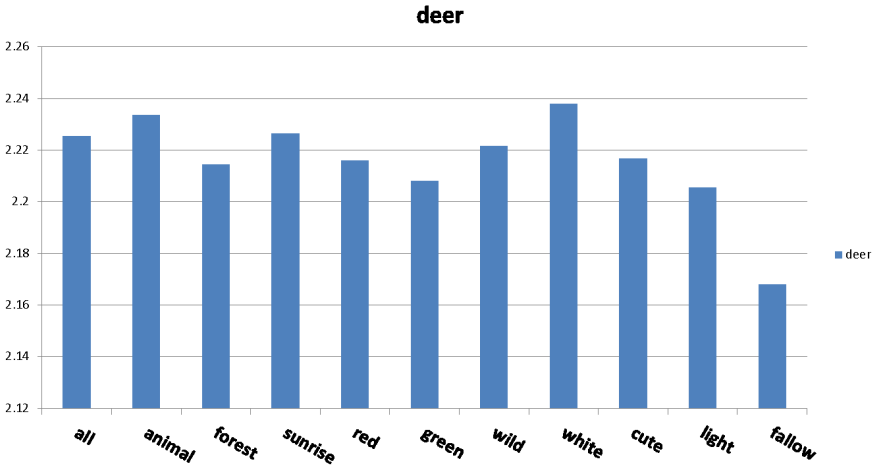

図 A.4: エントロピーの計算結果 (csift,deer)

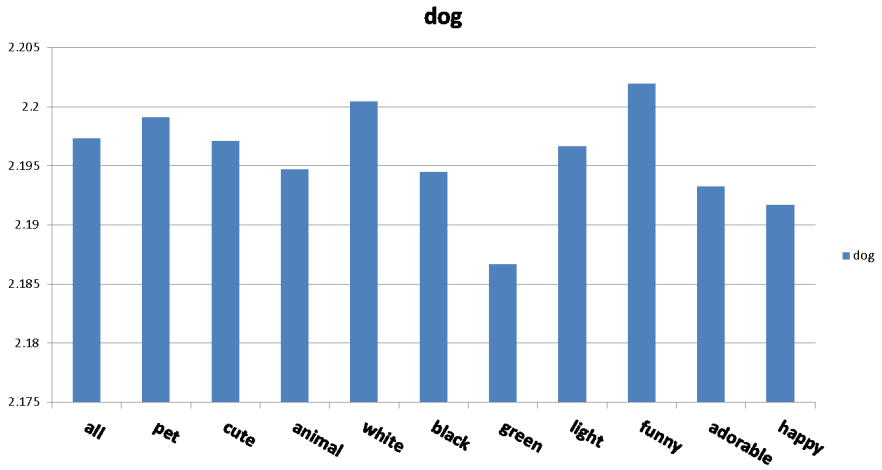

図 A.5: エントロピーの計算結果 (kmeans,dog)

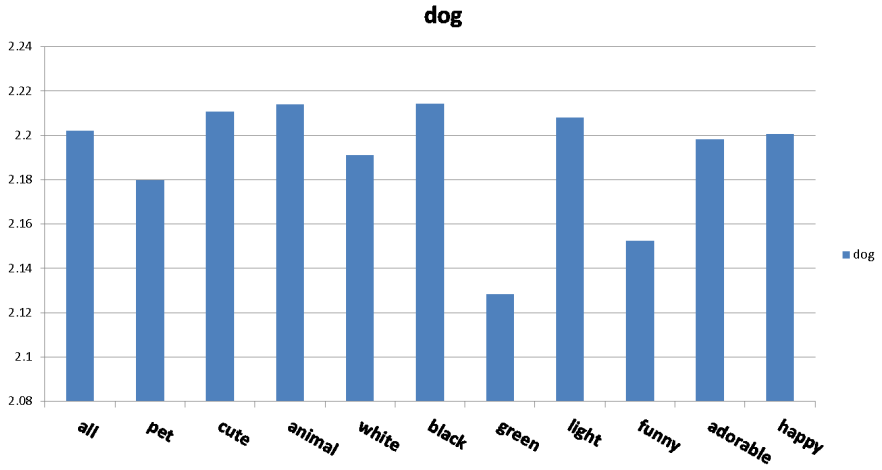

図 A.6: エントロピーの計算結果 (csift,dog)

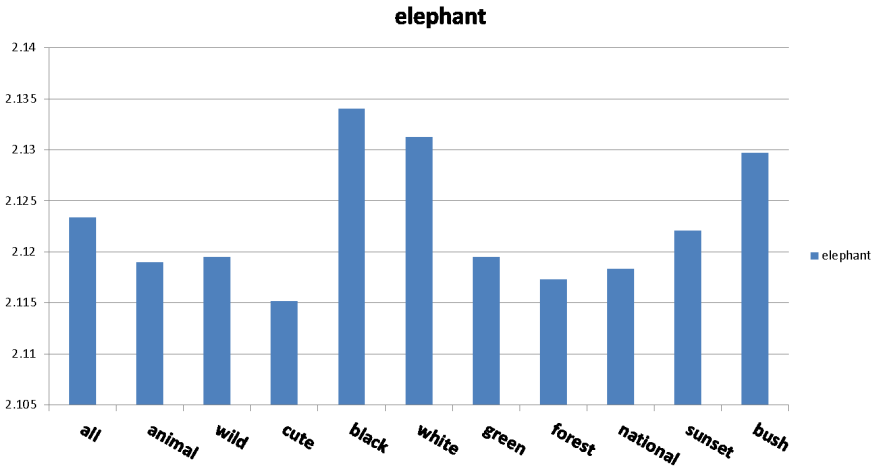

図 A.7: エントロピーの計算結果 (kmeans,elephant)

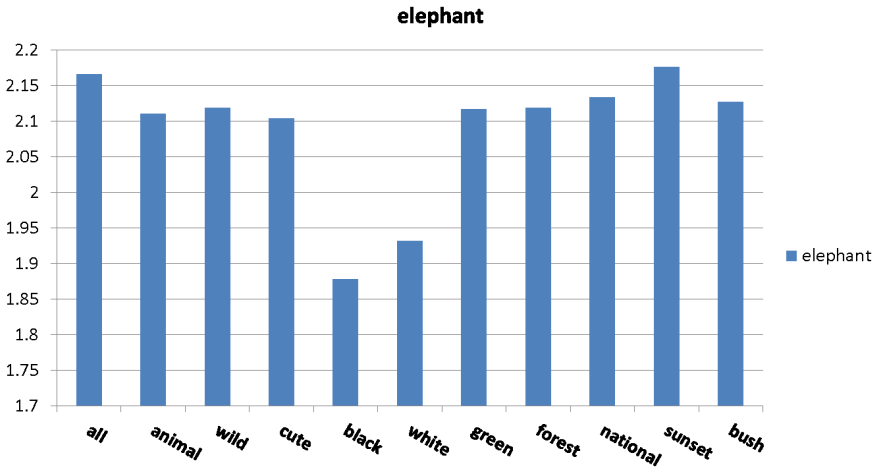

図 A.8: エントロピーの計算結果 (csift,elephant)

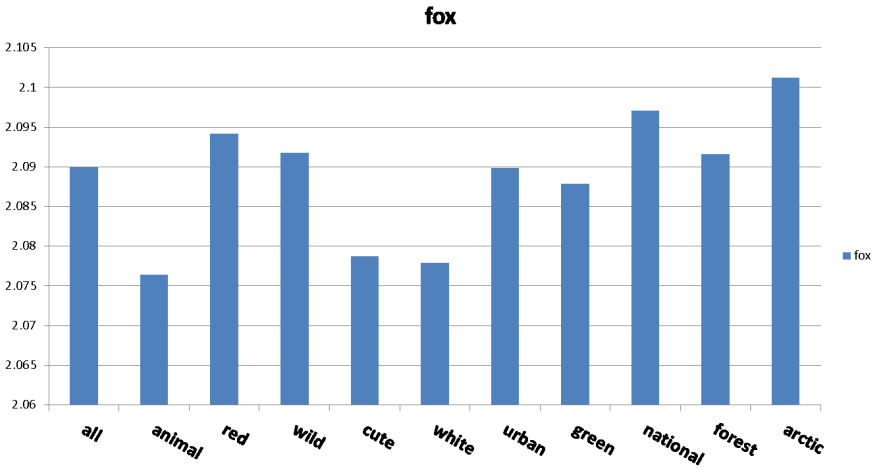

図 A.9: エントロピーの計算結果 (kmeans,fox)

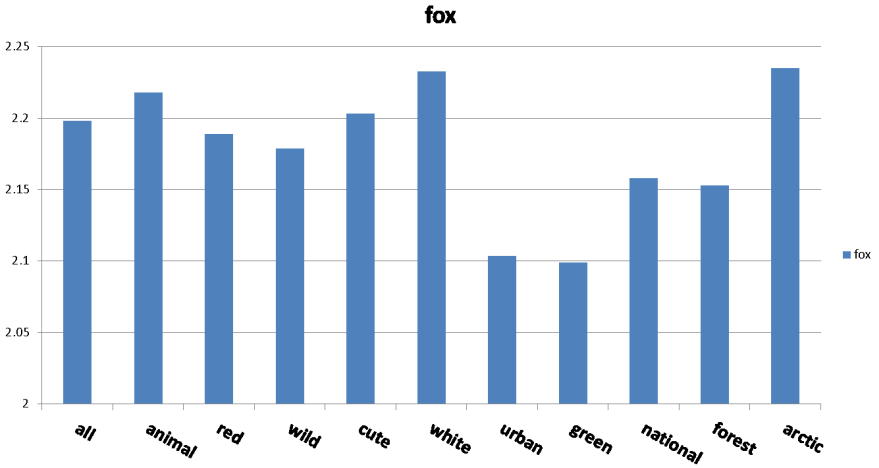

図 A.10: エントロピーの計算結果 (csift,fox)

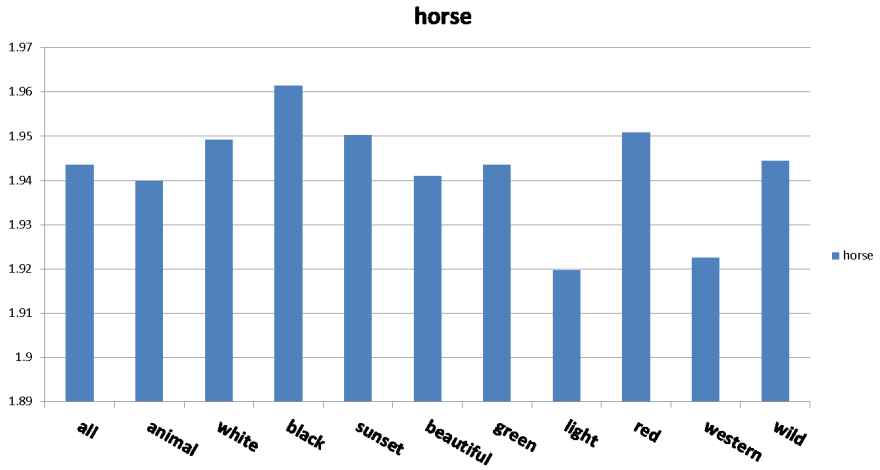

図 A.11: エントロピーの計算結果 (kmeans,horse)

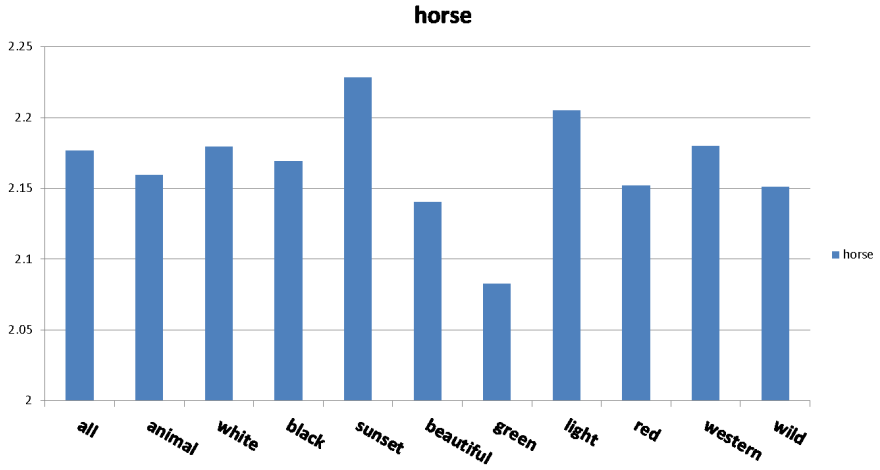

図 A.12: エントロピーの計算結果 (csift,horse)

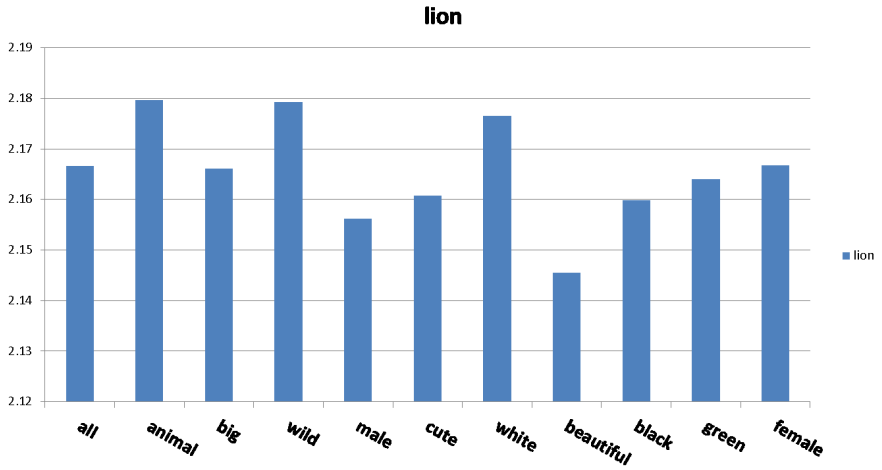

図 A.13: エントロピーの計算結果 (kmeans,lion)

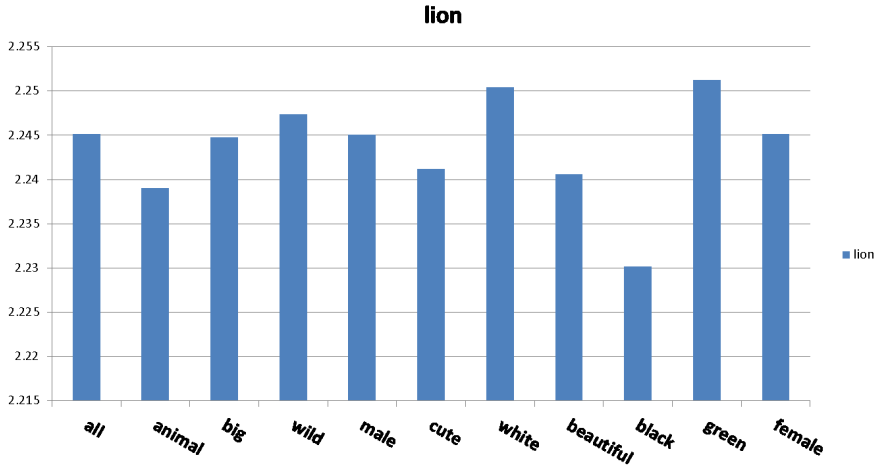

図 A.14: エントロピーの計算結果 (csift,lion)

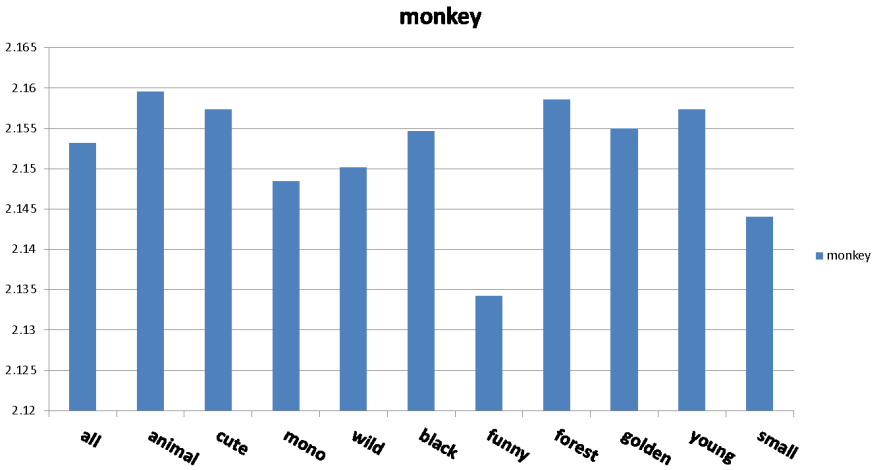

図 A.15: エントロピーの計算結果 (kmeans,monkey)

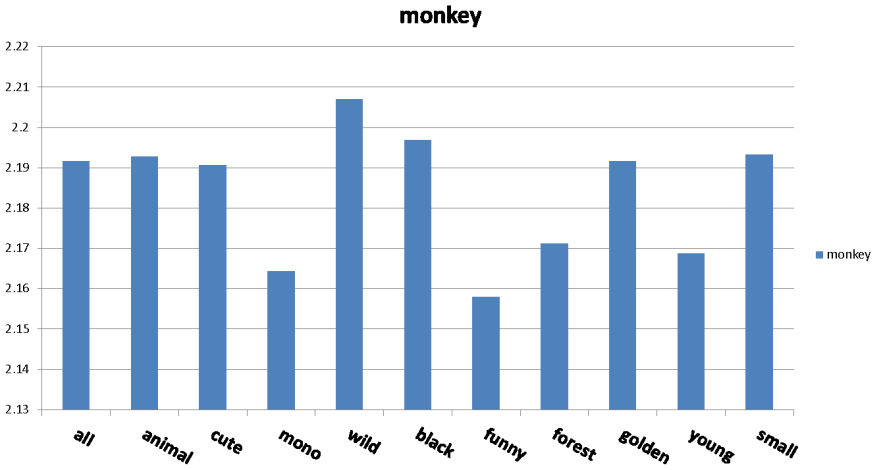

図 A.16: エントロピーの計算結果 (csift,monkey)

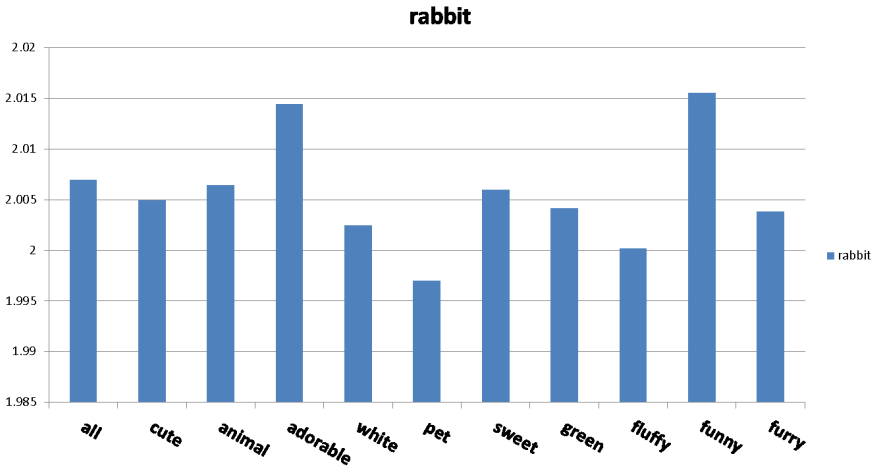

図 A.17: エントロピーの計算結果 (kmeans,rabbit)

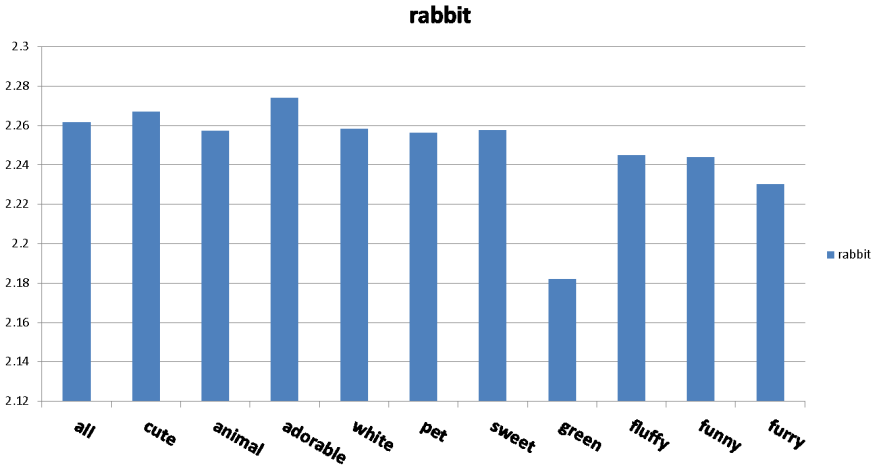

図 A.18: エントロピーの計算結果 (csift,rabbit)

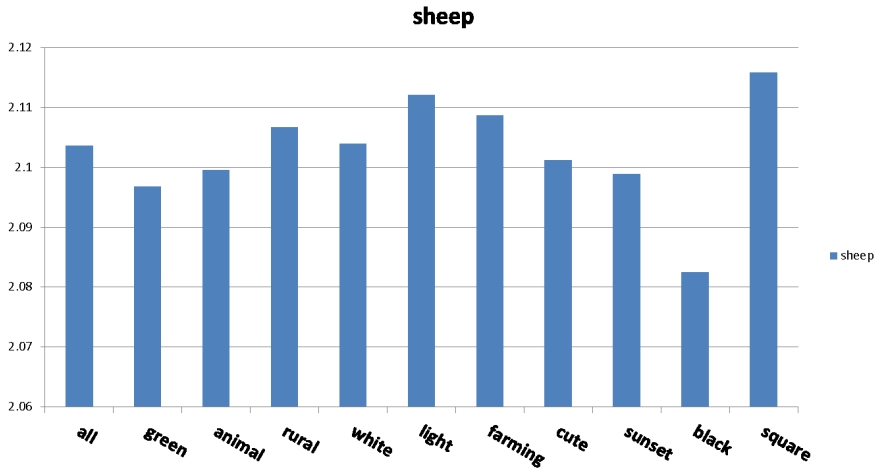

図 A.19: エントロピーの計算結果 (kmeans,sheep)

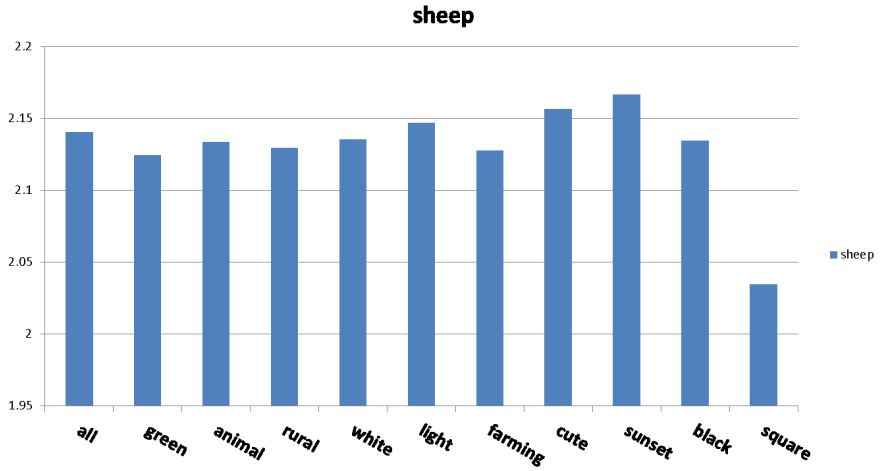

図 A.20: エントロピーの計算結果 (csift,sheep)